We did a var 1st approx with a factored dist — but we could do others.

E_q = -E_q log P - H_q

eg. Q: a tree structured dist?

Recall tree:

q(x_1, \ldots, x_N) = \prod_{i \in v} q_i \prod_{ij \in e} \frac{q_{ij}}{q_i q_j}

= \prod_{ij \in e} q_{ij} \frac{1}{\prod_{i \in v} q_i} \prod_{i \in v} q_i \frac{d_i - 1}{d_i}

so \ H_q = -\sum_{\text{values of } \mathbf{q}} \left[ \sum_{ij \in e} q_{ij} \log q_{ij} \right] + \sum_{\text{values}} \left[ \prod_{i \in v} q_i \log q_i \right]

tractable.
We can also compute

\[ -E_{Q} \log P \]

Recall:

\[
P(H|X) = \frac{1}{Z} \exp \left[ -\sum_{ij} \Theta_{ij}(H_{i}, H_{j}) - \sum_{i} \Theta_{i}(H_{i}) \right]
\]

\[
- \log P = \log Z + \sum_{ij} \Theta_{ij}(H_{i}, H_{j}) + \sum_{i} \Theta_{i}(H_{i})
\]

\[
\text{constant - not a prob, cause we're minimizing!}
\]

So

\[
- E_{Q} \log P = \log Z + \sum_{\text{pairs}} \left[ \sum_{ij \in \text{pairs}} q_{ij} t_{ij} \right] + \sum_{\text{values}} \left[ (d_{i} - 1) \sum_{i \in \text{values}} q_{i} t_{i} \right]
\]

\[
\text{x_i in other hand is notes; sorry!}
\]
Here I used the change of var
\[ t_{ij} = \Theta_{ij} + \Theta_i + \Theta_j \]
\[ t_i = \Theta_i \]

To get an expression that looks like entropy.

Now we want to min
\[ E_a \log Q - E_a \log P \]
for \( Q \) some fixed tree. (Which we chose)

Recall the \( q_{ij} \), \( q_i \), \( q_j \) are marginals

So we must
\[
\min \sum_{\text{values}} \left[ \sum_{i,j \in \mathcal{E}} q_{ij} [\log q_{ij} + t_{ij}] \right] \\
+ \sum_{\text{values}} \left[ \sum_i (d_i - 1) q_i [\log q_i + t_i] \right]
\]

\[ \text{st. } \sum_i q_{ij} = q_j, \quad \sum_j q_{ij} = q_i, \quad \sum q_i = 1 \]

(Cause these are marginals)
write

\[ \lambda_j \]

for LM's assoc. with

\[ \sum_j q_{ij} = q_i \]

\[ \sum_i q_{ij} = q_j \]

\[ \sum_i q_i = 1 \]

(Notice \( \lambda \) is a vector, \( \lambda_i \) is a scalar)

write lagrangian \( L \)

at stationary point

\[
\left[ \frac{\partial L}{\partial q_{ij}} \right] = 0 = \left[ \log q_{ij} \right]_{uv} + 1 + \left[ \delta_j^i \right] + \sum_i \left[ \lambda_i \right]_{uv} + \sum_j \left[ \lambda_j \right]_{uv} \]

so

\[
\left[ q_{ij} \right]_{uv} \propto \left[ e^{-\tilde{\tau}_{ij}} \right]_{uv} \left[ \Sigma \lambda_i \right]_{uv} \left[ \Sigma \lambda_j \right]_{uv} \]
compare with BP eqns

\[
\begin{align*}
\mathbf{q}_{ij} &\propto \mathbf{q}_{ij} \Phi_i \Phi_j \left[ \prod_{\text{all uic to } i} \mathbf{M}_i u \right] \left[ \prod_{\text{all uic to } j} \mathbf{M}_j v \right].
\end{align*}
\]

Conclusion

- Messages = log hi M's

Two outcomes:

1) We can fit a variational model of a single tree (MP as above)

2) loopy BP "like" fitting var m of tree without worrying about tree
Now what is happening in terms of M.P.? 

→ fix a tree. 

→ Interpret MP = convex hull of all states that can arise in this repn of G.M. 

→ We must have that 

Call the polytope that satisfies 

\[ \sum_{j} q_{ij} = q_i \quad \sum_{i} q_{ij} = q_j \quad \sum q_i = 1 \]

the Local Polytope = \(L_p\). 

→ Notice \(\text{MP} \subseteq L_p\).
Notice that for choice of tree $T$, a lot of $q_i = q_i q_j$ (cause the vars are independent and independent given parents).

For example:

Then in $T$, these two have $q_{ij} = q_i q_j$.

So we are finding the $q_{ij}, q_i, q_j$ that are in LP.

- meet indep constraints implied by $T$
- minimize $-E_q \log P + E_q \log Q$

(then extract info from $q$).
Extracting info from $Q$.

- If we're lucky, $q$ are integer.
  (might be a vert of $M_p$!)
  $\rightarrow$ nothing to do
- Else, it's a tree; $\rightarrow$ max product

Idea:

rather than

$$\min -E_a \log P + E_a \log Q$$

for $Q$ a tree,
do it for $Q \in L_P$.

$\rightarrow$ How do we get $E_a \log Q$?
Here is one strategy:

- Drop the tree
- Fit $q_i$ by using the expression for $E_{\log q}$ that came from the tree.

$$E_{\log q} = -\sum_{\text{edges}} \sum_{\text{pairs}} [q_{ij} \log q_{ij}] + \sum_{\text{verts}} \sum_{\text{values}} (d_i - 1) \frac{q_i}{\log q_i}$$

Notice I flipped order:

$$= H_{\infty}$$

This isn't the true exp. for $E_{\log q}$, but it's easy to eval.

A loopy b.p. = $\min E_{\alpha \log P} - H_{\alpha}$ s.t. $\alpha \in L.P.$
Notice the form of the cost function:

\[ E_\alpha \log p - H_\alpha. \]

Linear in \( \alpha \).

Some property of \( \alpha \) that approximates entropy.

Notice also that we're identifying points in LP (or MP) with probability distributions. It turns out that we can formalize this using the exponential family:

\[ p(x) = \exp \left[ \Theta^T \phi(x) - A(\Theta) \right] \]

Any p.d. that is written as
We will confine atten to case where:

- $\phi(x)$ are linearly independent
  (no real issue here, just creates a lot of ifs, ands, buts...)

- $\Theta$ is such that

$$A(\Theta) = \log \int \exp [\Theta^T \phi(x)] < \infty$$
Examples:

1D Normal Dist:

$$\exp \left[ \begin{pmatrix} x \\ \beta \end{pmatrix}^{T} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} (\alpha^2, \beta) - A(\Theta) \right]$$

here \( \alpha < 0 \); \( \text{std} = \frac{1}{2\alpha} \);

mean = \left( \frac{1}{2\alpha} \right)^2 \beta.

Multi-D ND:

follows easily.

Poisson dist:

recall this is dist on non-neg integers

\[ P(K) = \lambda^K \frac{e^{-\lambda}}{K!} \]
\[
p(k) = \exp \left[ \alpha \cdot k - A(\alpha) \right]
\]
\[
\alpha = \log \lambda, \text{ etc.}
\]

Discrete MRF:

\[
p(x) = \exp \left[ \Theta^T \left[ \begin{array}{c}
1(x_i) \\
\vdots \\
1(x_i, x_j) \\
\vdots
\end{array} \right] - A(\Theta) \right]
\]

\(\Pi_{\delta i}(x_i)\) is 1-hot vector
\(\Pi_{\delta j}(x_i, x_j)\) is 1-hot table, straightened into vector
$\Delta(\Theta)$ is extremely interesting

$$\nabla_{\Theta} A = \nabla_{\Theta} \left[ \log \int e^{\Theta \Phi} \, dx \right]$$

$$= \frac{1}{\int e^{\Theta \Phi} \, dx} \int \Phi e^{\Theta \Phi} \, dx$$

$$= E_{\Theta}[\Phi]$$

(recall—we've seen something like this before when talking about max-likelihood = max entropy.)

Now assume we have some $\Phi$ (likely indicator functions in our case).

We can define

$$\Lambda : \Theta \rightarrow M$$

$$\Lambda(\Theta) = E_{\Theta}[\Phi]$$

This is in $M$, cause $M$ is all possible expected,
Thus: \[ \Lambda \] is 1-1 (assuming \( \Phi \) are linearly indep).

(proof in Wainwright - mildly technical).

Now we want to consider dual of \( A(\Theta) \).

\[ A^*(\mu) = \sup_{\Theta \in \Theta} \left[ \langle \mu, \Theta \rangle - A(\Theta) \right] \]

Note this is a function of \( \mu \). Known as a conjugate dual.
Why is $A$ 1-1?

(sketched of proof - details in front-right)

$A(\theta)$ is convex

we must show for any $\mu$, there is some $\theta$ so that $E_{p(x;\theta)}[\phi] = \mu$.

But:

$E_{p(x;\theta)}[\phi] = \nabla_{\theta} A(\theta)$.

Under very mild conditions, map $x \to \frac{df}{dx}$ is 1-1 for $x$ convex.

- proof by drawing!
\( A(\Theta) \) is a convex function of \( \Theta \).

Recall:

\[
\frac{\partial A}{\partial \Theta_i} = e^{-A} \int e^{\Theta \varphi_i} \varphi_i \, dx
\]

So:

\[
\frac{\partial^2 A}{\partial \Theta_i \partial \Theta_j} = e^{-A} \int e^{\Theta \varphi_i} \varphi_i \varphi_j \, dx - \left[ e^{-A} \int e^{\Theta \varphi_i} \varphi_j \, dx \right] \left[ e^{-A} \int e^{\Theta \varphi_j} \varphi_i \, dx \right]
\]

\[
= E_p[\varphi_i \varphi_j] - E_p[\varphi_i] E_p[\varphi_j]
\]

\[
= \text{cov}(\varphi_i, \varphi_j)
\]

So:

\[
H_A = \text{cov mat} [\varphi]
\]

This is positive definite under our conditions (\( \varphi \) linearly indep.)
Now consider conjugate dual.

\[ f^*(\mu) = \sup_{\Theta} \left[ \langle \mu, \Theta \rangle - f(\Theta) \right] \]

for convex \( f \).

- assume \( f \) differentiable

then \( \mu = \nabla_{\Theta} f \)
Common visualization

- choose $\mu$ — what is $f^*(\mu)$?
- consider $\Theta$ such that

\[ \mu = \frac{df}{d\Theta} \]

The tangent line passes through $(\Theta, f(\Theta))$ and has slope $\mu$.

\[ y = \mu x + (f(\Theta) - \mu \Theta) \]

At $x = 0$, the tangent line at $y = -f(\mu)$.
for \( f \) convex, \( f^* \) is convex.

Show for \( f \in C^2 \), but generally true.

\[ \text{Proof: } f^*(p) = \sup_{x} [cp - f(x)] \]

\( f \) diff, convex so

\[ p = \frac{df}{dx} \text{ at sup.} \]

\( \frac{df}{dx} \) is a function, and \( \text{area} \) is \( 1-1 \)

So \( g \) s.t. \( g \circ df = 1 \) exists

So \( g(p) = x \). at sup

\[ g^*(p) = p \cdot g(p) - f(g(p)) \]

So \( f^*(p) = p \cdot g(p) - f(g(p)) \)

\[ \frac{df^*}{dp} = p \frac{dg}{dp} + g - f' \cdot \frac{dg}{dp} = g(p) \]

\[ \frac{d^2f^*}{dp^2} = \frac{dg}{dp} = \frac{dxc}{dx} \text{ but } \frac{dp}{dx} = \frac{ds}{dx^2} \]

so \( d^2f^* = \frac{1}{s(t)} > 0 \)
all this works in ND as well (ex: prove it!)

Thm (Fenchel-Moreau).

\[ f = (f^*)^* \]

iff

- \( f \) is proper, lower semi-continuous and convex
- \( f = \infty \)
- \( f = -\infty \)
Now consider $A(\theta) = \log Z(\theta)$, for an exp. dist.

1) $A^*(\mu) = \sup_{\theta} \langle \theta, \mu \rangle - A(\theta)$

is defined, convex.

2) for $\theta, \mu \in \mathcal{M}$, marginal polytope.

write $\Theta(\mu) = \Lambda^{-1}(\mu)$

then $A^*(\mu) = -H(p(x; \Theta(\mu)))$. 
Proof of 2 (sketch):

\[ \Lambda^{-1}(\mu) = \Theta \quad \text{such that} \quad E_{p(x;\Theta)}[\phi(x)] = \mu \]

(by defn of \( \Lambda \))

But if \( \mu = E_{p(x;\Theta)}[\phi] = \nabla_{\Theta} A(\Theta) \)

then \( \Theta \) is \( \text{sup} \)

so

\[ -H(p(x;\Theta(\mu))) = E_{p(x;\Theta(\mu))} \left[ \langle \Theta, \phi(x) \rangle - A(\Theta) \right] \]

\[ = \langle \Theta, \mu \rangle - A(\Theta) \]

\[ = A^*(\mu) \quad (\text{cause } \Theta \text{ is sup}) \]
\[ A(\Theta) = \sup_{\mu \in M} \{ \langle \Theta, \mu \rangle - A^*(\mu) \} \]

(A is lower semicontinuous - see notes; then Fenchel-Moreau means

\[ (A^*)^* = A \]

and \( (A^*)^*(\Theta) = \sup_{\mu \in M} \{ \langle \Theta, \mu \rangle - A^*(\mu) \} \)

\underline{compare: with}

\[ E_Q = -E_{\Theta \log p} + E_Q \log q \]

which we minimized to build von model

\[ E_Q \log p \rightarrow \langle \Theta, \mu \rangle \]
Now we have

\[ \sup_{\mu \in \mathcal{M}} \left\{ <\theta, \mu> - A^*(\mu) \right\} \rightarrow \circ \]

is attained at

\[ \mu = E_{p(x; \theta)} [\phi] \]

so solving \( \circ \) gives

- log partition function
- set of mean pars
  (for our purposes)
  \[ \arg \max \]

BUT

\[ M \text{ is hard.} \]

\[ A^*(\mu) \text{ is hard.} \]
Now we can unify alg.

Mean field, single tree, etc

for any $\mu \in \mathbb{M}$,

$$A(\Theta) \geq \langle \mu, \Theta \rangle - \hat{A}^*(\mu).$$

Now consider $T \subset \mathbb{M}$

$$\sup_{T} \left\{ \langle \Theta, \mu \rangle - \hat{A}^*(\mu) \right\} = A_{\text{MF}}(\Theta)$$

This norm is the same as our exp

but $\Theta - \text{sign of max}$

must have $A_{\text{MF}}(\Theta) \leq A(\Theta)$.
1) recall for a tree structured model

\[ H(q) = -\sum_{x \text{ values}} \left[ \sum_{\text{events}} q_i(x_i) \log q_i(x_i) \right] \]

\[ - \sum_{\text{i,j edges e values}} \left[ \sum_{x_i, x_j} q_{ij}(x_i, x_j) \log \frac{q_{ij}(x_i, x_j)}{q_i(x_i) q_j(x_j)} \right] \]

2) approximate

\[ \hat{A}^*(\mu) \approx -H(\mu) \quad \text{computed using tree expression} \]

Bethe approx

3) \# M

local polytope, consistency constraints for pairwise marginals
4) Solve
\[ \sup_{\mu \in \mathcal{Z}} \{ <\theta, \mu > + H(\mu) \} = A_{\text{LB}}(\theta) \]

we must have
\[ A_{\text{LB}}(\theta) > A(\theta) \]

But we can now explore other approximations:
- e.g. insert constraints so that

\[ L \subset L \subset M \]

defined by marginal constraints

marginal + extra

true marginal polytope
Here's one construction.

Assume some vector \( \mu \), which might be in \( M \).

Construct the matrix:

\[
M = \begin{bmatrix}
\mu_1 & \mu_2 & \cdots \\
\mu_1^2 & \mu_1 \mu_2 & \cdots \\
\vdots & \vdots & \ddots \\
\mu_1^2 & \mu_1 \mu_2 & \cdots
\end{bmatrix}
\]

\( M \) is a covariance matrix. So \( M \succ 0 \).